# A FAST ALGORITHM FOR STOCHASTIC MATCHING WITH APPLICATION TO ROBUST SPEAKER VERIFICATION

*Qi Li, S. Parthasarathy†, and Aaron E. Rosenberg†*

Speech Research Department
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
Email: qli@bell-labs.com

†Speech & Image Processing Services Research
AT&T Labs
Murray Hill, NJ 07974
Email: sps,aer@research.att.com

## ABSTRACT

Acoustic mismatch between training and test environments is one of the major problems in telephone-based speaker recognition. Speaker recognition performances are degraded when an HMM trained under one set of conditions is used to evaluate data collected from different telephone channels, microphones, etc. The mismatch can be approximated as a linear transform in a cepstral domain. In this paper, we present a fast, efficient algorithm to estimate the parameters of the linear transform for real-time applications. Using the algorithm, test data are transformed toward the training conditions by rotation, scale, and translation without destroying the the detailed characteristics of speech, then, speaker dependent HMM's can be used to evaluate the details under the same condition as training. Compared to cepstral mean subtraction (CMS) and other bias removal techniques, the proposed linear transform is more general since CMS and others only consider translation; compared to maximum-likelihood approaches for stochastic matching, the proposed algorithm is simpler and faster since iterative techniques are not required. The proposed algorithm improves the performance of a speaker verification system in the experiments reported in this paper.

## 1. INTRODUCTION

For speaker recognition, a speaker-dependent hidden Markov model (HMM) for a true speaker is usually trained based on training data collected in one enrollment session. The HMM, therefore, matches the probability density function (*pdf*) of the training data perfectly. In a verification session, test data are very often collected through a different telephone channel and handset. Since the acoustic condition is different from the enrollment session, it usually causes a mismatch between the test data and the trained HMM. Speaker recognition performance is degraded by the mismatch.

The mismatch can be represented as a linear transform in the cepstral domain:

$$y = \mathbf{A}x + b, \tag{1}$$

where $x$ is a vector of the cepstral frame of a test utterance; $\mathbf{A}$ and $b$ are the matrix and vector which need to be estimated for every test utterance; and $y$ is a transformed vector. Geometrically, $b$ represents a translation and $\mathbf{A}$ represents both scale and rotation. When $\mathbf{A}$ is diagonal, it is only a scaling operation. An analysis of the reason for using a linear transformation is beyond the scope of this paper. Interested readers are referred to [1].

Cepstral mean subtraction (CMS) is a fast, efficient technique for handling mismatch in both speaker and speech recognition. It estimates $b$ and assumes $\mathbf{A}$ to be an identity matrix. In [2], the vector $b$ was estimated by long term average, short term average, and a maximum likelihood approach. In [3, 4], maximum likelihood (ML) approaches were used to estimate $b$, a diagonal $\mathbf{A}$, and model parameters for HMM's for stochastic matching. Very recently, a least-squares solution of the linear transform parameters was briefly introduced in [1].

In this paper, we consider a general linear transform, i.e. $\mathbf{A}$ is a full matrix, and $b$ is a vector. The approach is to have the overall distribution of test data match the overall distribution of training data. Then a speaker dependent (SD) HMM trained on the training data is applied to evaluate the details of the test data. This is based on the assumption that differences between speakers are mainly on the details which have been characterized by HMM's. The associated fast algorithm for real-time applications is also given in this paper. Compared to CMS and other bias removal techniques [2, 5], the proposed linear transform is more general since CMS and others only consider the translation; compared to the ML approaches [2, 5, 3, 4], the algorithm is simpler and faster since iterative tech-

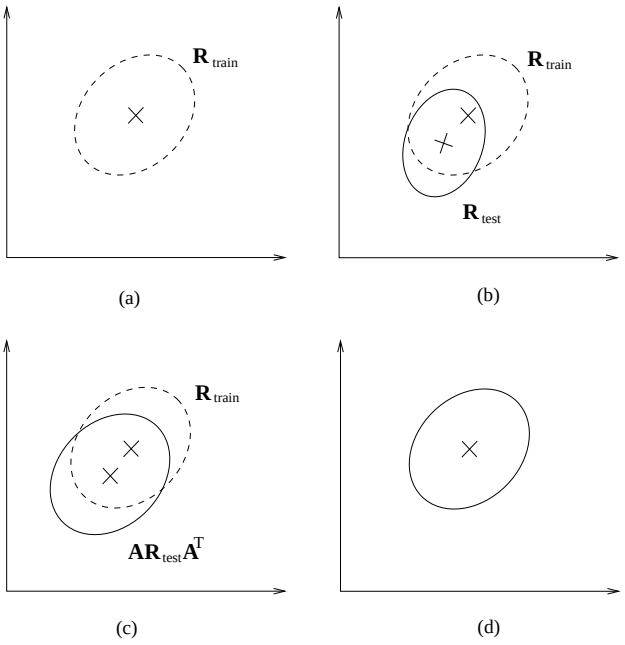


Figure 1: A geometric interpretation of the fast stochastic matching.

niques are not required and the estimation of the linear transform parameters is separated from the HMM training and test.

## 2. THE CONCEPT OF THE FAST STOCHASTIC MATCHING

We use Fig. 1 as a geometric interpretation of the proposed matching algorithm. In Fig. 1 (a), the dashed line is a contour of training data. In Fig. 1 (b), the solid line is a contour of test data. Due to different channels, noise levels, and telephone transducers, the mean of the test data is translated from the training data; the distribution is shrunk [6] and rotated from the HMM training condition. The mismatch may cause a wrong decision when using the trained HMM to score the mismatched test data. By applying the proposed algorithm, we first find a covariance matrix, $\mathbf{R}_{train}$, from the training data which characterizes the overall distribution approximately. Then, we find a covariance matrix, $\mathbf{R}_{test}$, from the test data and estimate the parameters of the $\mathbf{A}$ matrix for the linear transform in (1). After applying the first transform, the overall distribution of the test data is scaled and rotated, $\mathbf{A}\mathbf{R}_{test}\mathbf{A}^T$, to be same as the training data except for the difference of the means, as shown in Fig. 1 (c). In the second step, we find the difference of the means, and translate the test data to the same location of the training data as shown in Fig. 1 (d), where the contour of the transformed test data is overlapped to the contour of the training data.

If the test data from a true speaker mismatch the HMM training condition, the data will be transformed to match the trained HMM approximately. If the test data from a true speaker match the training condition, the calculated $\mathbf{A}$ and $b$ are close to an identity matrix and a zero vector respectively, so the transform will not effect the HMM scores much.

This technique attempts to improve mismatch whether the mismatch occurs because test and training conditions differ or because the test and training data originate from different speakers. It is reasonable to suppose that speaker characteristics are found mainly in the details of the representation. However, to the extent that they are also found in global features, this technique would increase the matching scores between true speaker models and imposter test utterances. Performance, then, could possibly degrade particularly when other sources of mismatch are absent, that is, when test and training conditions are actually matched. However, experiments in this paper will show that performances overall do improve.

## 3. FAST ESTIMATION FOR A GENERAL LINEAR TRANSFORM

In a speaker verification training session, we collect multiple utterances with the same content, and use a covariance matrix $\mathbf{R}_{train}$, a mean vector $m_{train}$ to represent the overall distribution of the training data of all the training utterances in a cepstral domain. They are defined as follows:

$$\mathbf{R}_{train} = \frac{1}{U}\sum_{i=1}^{U}\frac{1}{N_i}\sum_{j=1}^{N_i}(x_{i,j} - m_i)(x_{i,j} - m_i)^T, \quad (2)$$

and

$$m_{train} = \frac{1}{U}\sum_{i=1}^{U} m_i, \quad (3)$$

where $x_{i,j}$ is the $j$th non-silence frame in the $i$th training utterance, $U$ is the total number of training utterances, $N_i$ and $m_i$ are the total number of non-silence frames and the mean vector of the $i$th training utterance respectively, and $m_{train}$ is the average mean vector of the non-silence frames of all training utterances.

In a test session, only one utterance will be collected and verified at a time. The covariance matrix for the test data is

$$\mathbf{R}_{test} = \frac{1}{N_f}\sum_{j=1}^{N_f}(y_j - m_{test})(y_j - m_{test})^T, \quad (4)$$

where $y_j$ and $m_{test}$ are a non-silence frame and the mean vector of the test data, $N_f$ is the total number of non-silence frames.

The proposed criterion for parameter estimation is to have $\mathbf{R}_{test}$ match $\mathbf{R}_{train}$ through a rotation, scale, and translation (RST) of the test data. For rotation and scale, we have the following equation.

$$\mathbf{R}_{train} - \mathbf{A}\mathbf{R}_{test}\mathbf{A}^T = 0, \tag{5}$$

where $\mathbf{A}$ is defined as in (1); $\mathbf{R}_{train}$ and $\mathbf{R}_{test}$ are defined as in (2) and (4). By solving (5), we have the $\mathbf{A}$ matrix for (1),

$$\mathbf{A} = \mathbf{R}_{train}^{\frac{1}{2}}\mathbf{R}_{test}^{-\frac{1}{2}}. \tag{6}$$

Then, the translation term $b$ of (1) can be obtained by

$$b = m_{train} - m_{rs} = m_{train} - \frac{1}{N_f}\sum_{j=1}^{N_f}\mathbf{A}x_j \tag{7}$$

where $m_{train}$ is defined as in (3); $m_{rs}$ is a mean vector of rotated and scaled frames; $N_f$ is the total number of non-silence frames of a test utterance; $x_j$ is the $j$th cepstral vector frame.

To verify a given test utterance against a set of true speaker's models (consisting of a SD HMM plus $\mathbf{R}_{train}$, $m_{train}$), first $\mathbf{R}_{test}$, $\mathbf{A}$ and $b$ are calculated by using (4), (6), and (7), then all test frames are transformed by (1) to reduce the mismatch.

## 4. GENERAL PHRASE SPEAKER VERIFICATION

The above stochastic matching algorithm has been applied to a text-dependent speaker verification system using general phrase passwords. The system was described by Parthasarathy and Rosenberg in [7]. Stochastic matching is included in the front-end processing to further improve the system robustness and performance.

The system block diagram with stochastic matching is shown in Fig. 2. After a speaker claims an identity (ID), the system expects the same phrase obtained in the associated training session. First, a speaker independent (SI) phone recognizer segments the input utterance into a sequence of phones by forced decoding using the transcription saved from the enrollment session. Since the SD models are trained on a small amount of data from a single session, they can't be used to provide a reliable and consistent phone segmentations. So the SI phone models are used. On the other hand, the cepstral coefficients of the utterance from the test speaker is transformed to match the

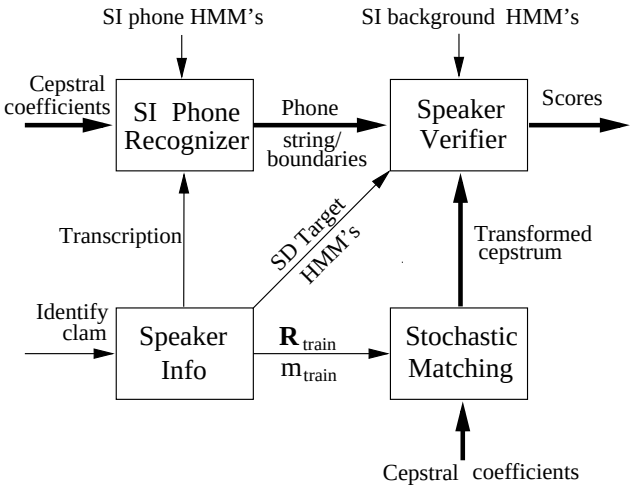


Figure 2: A phrase-based speaker verification system with stochastic matching

training data distribution by computing Eqs. (4), (6), (7), and (1). Then, the transformed cepstral coefficients, decoded phone sequence, and associated phone boundaries are transmitted to a verifier. In the verifier, a log-likelihood-ratio score is calculated based on the log-likelihood scores of target and background models.

$$L_R(\mathbf{O};\Lambda_t;\Lambda_b) = L(\mathbf{O},\Lambda_t) - L(\mathbf{O},\Lambda_b) \tag{8}$$

where $\mathbf{O}$ is the observation sequence over the whole phrase, and $\Lambda_t$ and $\Lambda_b$ are the target and background models respectively. The background model is a set of HMM's for phones. The target model is one HMM with multiple states for whole phrase. As reported in [7], this configuration provides the best results in experiments. Furthermore,

$$L(\mathbf{O},\Lambda_t) = \frac{1}{N_f}P(\mathbf{O}|\Lambda_t), \tag{9}$$

where $P(\mathbf{O}|\Lambda_t)$ is the log-likelihood of the phrase evaluated by one HMM, $\Lambda_t$, using Viterbi decoding, and $N_f$ is the total number of non-silence frames in the phrase.

$$L(\mathbf{O},\Lambda_b) = \frac{1}{N_f}\sum_{i=1}^{N_p}P(\mathbf{O}_i|\Lambda_{b_i}) \tag{10}$$

where $P(\mathbf{O}_i|\Lambda_{b_i})$ is the log-likelihood of the $i$th phone, $\mathbf{O}_i$ is the segmented observation sequence over the $i$th phone, $\Lambda_{b_i}$ is an HMM for the $i$th phone, $N_p$ is the total number of the decoded non-silence phones, and $N_f$ is the same as above.

A finial decision on rejection or acceptance is made based on the $L_R$ score with a threshold. If a significantly different phrase is given, the phrase could be rejected by the SI phone recognizer before using the verifier.

## 5. FEATURES AND DATABASE

The feature vector in this paper is composed of 12 cepstrum and 12 delta cepstrum coefficients. The cepstrum is derived from a 10th order LPC analysis over a 30 ms window. The feature vectors are updated at 10 ms intervals.

The experimental database consists of fixed phrase utterances recorded over the long distance telephone networks by 100 speakers, 51 male and 49 female. The fixed phrase, common to all speakers, is "I pledge allegiance to the flag" with an average length of 2 seconds. Five utterances of each speaker recorded in one session are used to train a SD HMM plus $\mathbf{R}_{train}$, $m_{train}$ for the linear transform. For testing, we used 50 utterances recorded from a true speaker at different sessions (different telephone channels at different times), and 200 utterances recorded from 50 impostors of the same gender at different sessions. For model adaptation, the second, fourth, sixth, and eighth test utterances from the tested true speaker are used to update the associated HMM plus $\mathbf{R}_{train}$, $m_{train}$ for verifying succeeding test utterances.

The target models for phrases are left-to-right HMM's. The number of the states are 1.5 times the total number of phones in the phrases. There are 4 Gaussian components associated with each state. The background models are concatenated phone HMM's trained on a telephone speech database from different speakers and texts. Each phone HMM has 3 states with 32 Gaussian components associated with each state.

Due to unreliable variance estimates from limited amount of training data, a global variance estimate is used as a common variance to all Gaussian components [7] in the target models.

## 6. EXPERIMENTAL RESULTS

The experimental results are listed in Table 1. These are the averages of individual equal-error rates over the 100 evaluation speakers. The baseline results are obtained with log-likelihood-ratio scores using phrase-based target model and phone-based speaker background models. The equal-error rates (EER's) without and with adaptation are 5.98% and 3.94% respectively. When using CMS, the EER's are 3.03% and 1.96%. When using the proposed algorithm (RST), the equal error rates are 2.61% and 1.80%.

Table 1: **Experimental Results in Average Equal-Error Rates (%)**

| Algorithms | No Adaptation | With Adaptation |
|---|---|---|
| Baseline | 5.98 | 3.94 |
| CMS | 3.03 | 1.96 |
| RST(proposed) | 2.61 | 1.80 |

## 7. CONCLUSIONS

A simple, fast and efficient algorithm for stochastic matching has been presented. The algorithm is applied to a general phrase speaker verification system. In the experiments, when there is no model adaptation, the proposed algorithm improves equal-error rates by 56% compared with a baseline system without any stochastic matching, and 14% compared with a system using CMS. When model adaptation is applied, the improvements are 54% and 8%. Less improvement is obtained because the SD models are updated to fit different acoustic conditions. The proposed algorithm can also be applied to speaker identification and other applications to improve system robustness.

## 8. REFERENCES


[1] R. J. Mammone, X. Zhang, and R. P. Pamachandran, "Robust speaker recognition," *IEEE Signal Processing Magzine*, vol. 13, pp. 58–71, Sept. 1996.

[2] A. E. Rosenberg, C. H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for hmm-based speaker verification," in *Proceedings of ICSLP-94*, pp. 1835–1838, 1994.

[3] A. Sankar and C. H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 190–202, May 1996.

[4] A. C. Surendran, *Maximum-likelihood stochastic matching approach to non-linear equalization for robust speech recognition*. PhD thesis, Rutgers University, Busch, NJ, May 1996.

[5] M. G. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 19–30, Jan. 1996.

[6] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1659–1671, Nov. 1989.

[7] S. Parthasarathy and A. E. Rosenberg, "General phrase speaker verification using sub-word background models and likelihood-ratio scoring," in *Proceedings of ICSPL-96*, Oct. 1996.